Testing Network Hypotheses

Steve Borgatti

Revised 20 July, 2004. Colchester, U.K.

Units of Analysis

- Dyadic (tie-level)
  - The raw data
  - Cases are pairs of actors
  - Variables are attributes of the relationship among pairs (e.g., strength of friendship; whether give advice to; hates)
  - Each variable is an actor-by-actor matrix of values, one for each pair
- Monadic (actor-level)
  - Cases are actors
  - Variables are aggregations that count number of ties a node has, or sum of distances to others (e.g., centrality)
  - Each variable is a vector of values, one for each actor
- Network (group-level)
  - Cases are whole groups of actors along with ties among them
  - Variables aggregations that count such things as number of ties in the network, average distance, extent of centralization, average centrality
  - Each variable has one value per network

Types of Hypotheses

- Dyadic (multiplexity)
  - Friendship ties lead to business ties
  - Social ties between leads to less formal contractual ties (embeddedness)
- Monadic
  - Actors with more ties are more successful (social capital)
- Network
  - Teams with greater density of communication ties perform better (group social capital)
- Mixed Dyadic-Monadic (autocorrelation)
  - People prefer to make friends (dyad level) with people of the same gender (actor level) (homophily)
  - Friends influence each other’s opinions

Statistical Issues

- Samples non-random
- Often work with populations
- Observations not independent
- Distributions unknown

Solutions

- Non-independence
  - Model the non-independence explicitly as in HLM
    - Assumes you know all sources of dependence
  - Permutation tests
- Non-random samples/populations
  - Permutation tests

Logic of Permutation Test

- Compute test statistic
  - e.g., correlation or difference in means
  - Correlation between centrality and salary is 0.384 or difference in mean centrality between the boys and the girls is 4.95.
  - Ask what are the chances of getting such a large correlation or such a large difference in means if the variables are actually completely independent?
- Wait! If the variables are independent, why would the correlation or difference in means be anything but zero?
  - Sampling
  - "Combinatorial chance": if you flip coin 10 times, you expect 5 heads and 5 tails, but what you actually get could be quite different

Logic of Permutation Test

- So to evaluate an observed correlation between two variables of 0.384, we want to
  - correlate thousands of variables similar to the ones we are testing that we <u>know</u> are truly independent of each other, and
  - see how often these independent variables are correlated at a level as large as 0.384
    - The proportion of random correlations as large observed value is the p-value of the test
- How to obtain thousands of independent variables whose values are assigned independently of each other?
  - Fill them with random values
    - But need to match distribution of values
  - Permute values of one with respect to the other

Outline of Permutation Test

- Get observed test statistic
- Construct a distribution of test statistics under null hypothesis
  - Thousands of permutations of actual data
- Count proportion of statistics on permuted data that are as large as the observed
  - This is the p-value of the test

Monadic Hypotheses

	Centrality	Grades
bill	10	2.1
maria	20	9.5
mikko	40	7.3
esteban	30	4.1
jean	70	8.1
ulrik	50	8.1
joao	40	6.6
myeong-gu	50	3.3
akiro	60	9.1
chelsea	10	7.2

Dyadic Hypotheses

- Hubert / Mantel QAP test
  - All variables are actor-by-actor matrices
  - We use one relation (dyadic variable) to predict another
  - Test statistic is $\gamma = \sum_i \sum_j x_{ij} y_{ij}$
  - Significance is

$$prop(\gamma \geq \gamma^P),$$

$$\gamma^P = \sum_i \sum_j x_{ij} y_{p(i)p(j)}$$

- QAP correlation & MR-QAP multiple regression

Friendship (X)

	Jim	Jill	Jen	Joe
Jim	-	1	0	1
Jill	1	-	1	0
Jen	0	1	-	1
Joe	1	0	1	-

Proximity (Y)

	Jim	Jill	Jen	Joe
Jim	-	3	9	2
Jill	3	-	1	15
Jen	9	1	-	3
Joe	2	15	3	-

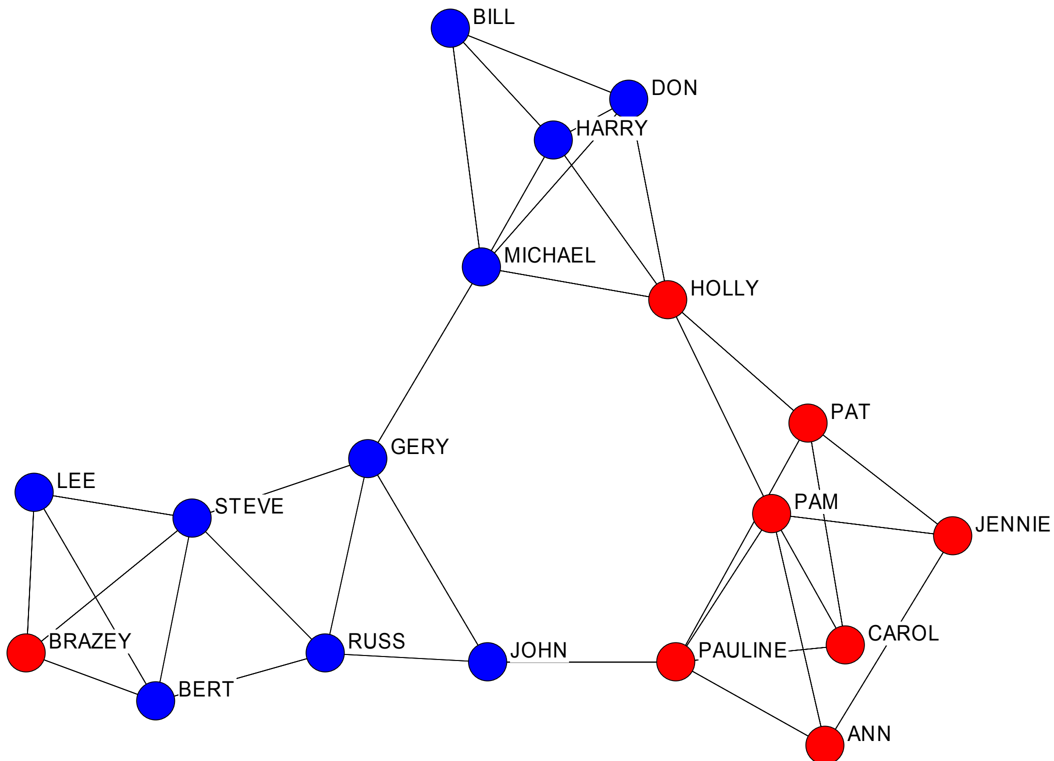
Dyadic/Monadic Hypotheses

- One dyadic (relational) variable, one monadic (actor attribute) variable
  - Technically known as autocorrelation
- Diffusion
  - adjacency leads to similarity in actor attribute
    - Spread of information; diseases
- Selection
  - similarity leads to adjacency
    - Homophily: birds of feather flocking together
    - Heterophily: disassortative mating
- Tom Snijders' SIENA model

Categorical Autocorrelation

- Nodes partitioned into mutually exclusive categories, e.g., gender or race
- We expect more ties within group than between
  - Boys interact w/ boys, girls w/ girls
- Count up number of ties between all ordered pairs of groups:
  - boys to boys, boys to girls, girls to boys, girls to girls
- Compare with number expected given independence of interaction and node characteristic
  - i.e., if people choose partners without regard for gender

Campnet Example



Observed

	Female	Male
Female	12	7
Male	7	16

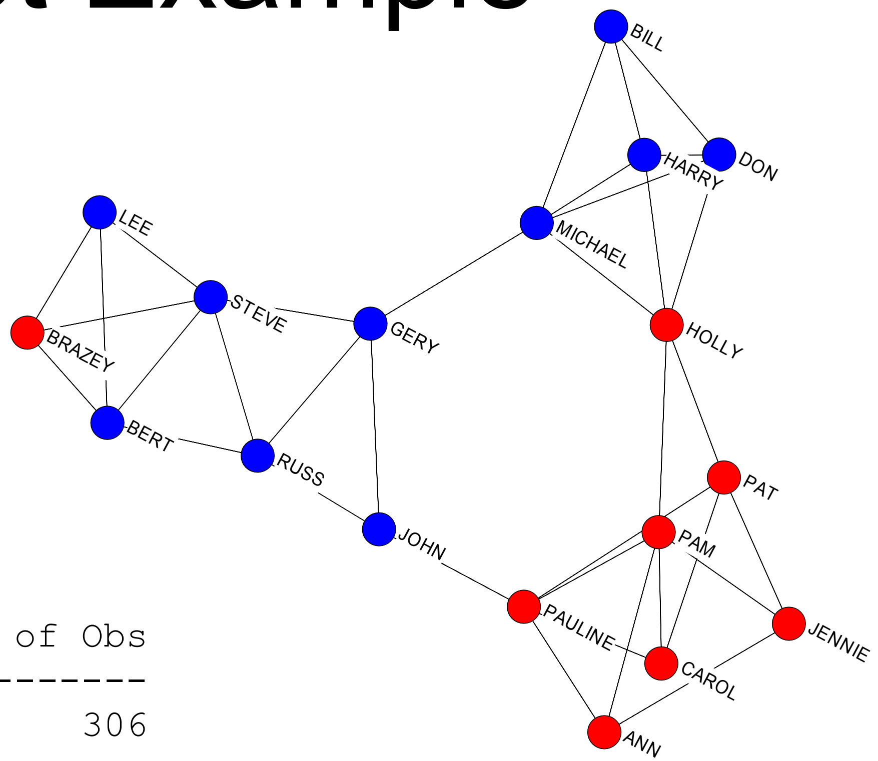
Expected

	Female	Male
Female	6.4	18.3
Male	18.3	10.3

Ratio

	Female	Male
Female	1.87	0.38
Male	0.38	1.55

Campnet Example




```
Density Table

               1     2
           Femal  Male
           ----- -----
  1 Fem    0.429 0.087
  2 Mal    0.087 0.356

MODEL FIT

R-square Adj R-Sqr Probability    # of Obs
-------- --------- ----------- -----------
   0.127     0.124       0.001         306

REGRESSION COEFFICIENTS

                Un-stdized     Stdized                 Proportion  Proportion
 Independent Coefficient Coefficient Significance    As Large    As Small
 ----------- ----------- ----------- ------------ ----------- -----------
   Intercept    0.087500    0.000000        1.000       1.000       0.001
     Group 1    0.341071    0.313982        0.001       0.001       0.999
     Group 2    0.268056    0.290782        0.001       0.001       0.999
```

Continuous Autocorrelation

- Each node has score on continuous variable, such as age or rank
- Positive autocorrelation exists when nodes of similar age tend to be adjacent
  - Friendships tend to be homophilous wrt age
  - Mentoring tends to be heterophilous wrt age
- Can measure similarity via difference or product

Geary's C

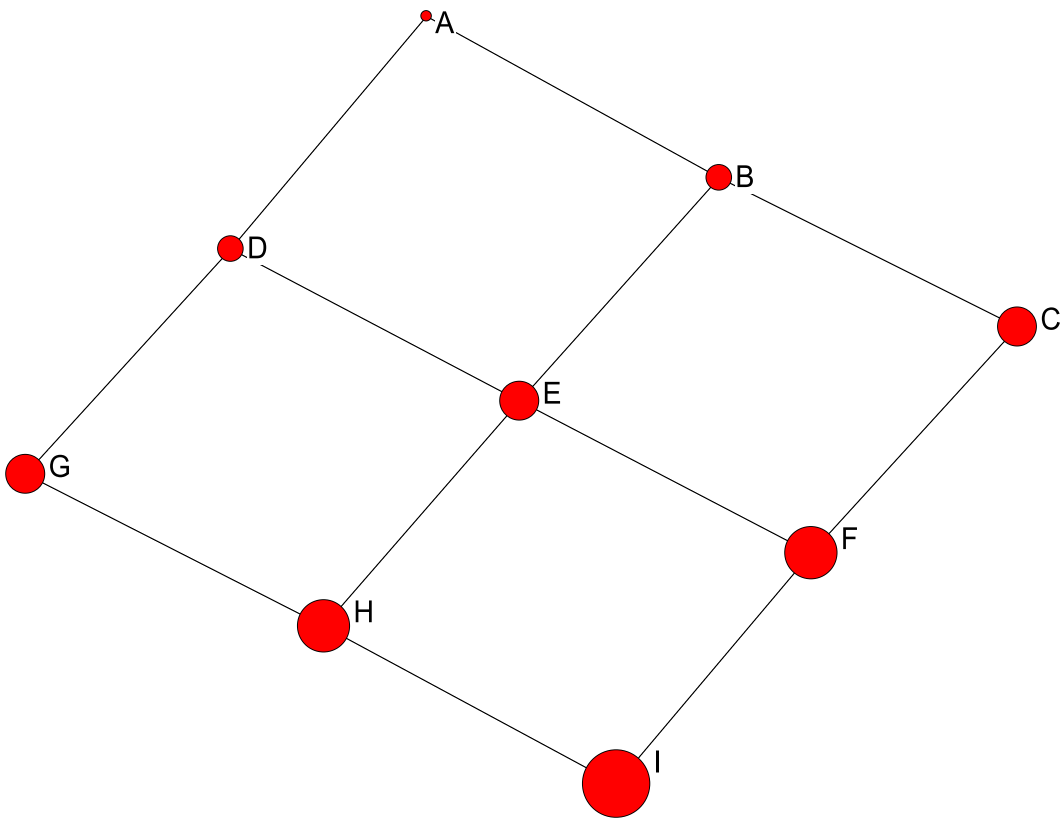
- Let Wij > 0 indicate adjacency of nodes i and j, and Xi indicate the score of node i on attribute X (e.g., age)

$$C = (n-1)\frac{\sum_i \sum_j w_{ij}(x_i - x_j)^2}{2\sum_{i,j} w_{ij} \sum_i (x_i - \overline{x})^2}$$

- Range of values: 0 <= C <= 2
  - C=1 indicates independence;
  - C > 1 indicates negative autocorrelation;
  - C < 1 indicates positive autocorrelation (homophily)

Positive Autocorrelation

(Similars adjacent; Geary's C < 1)

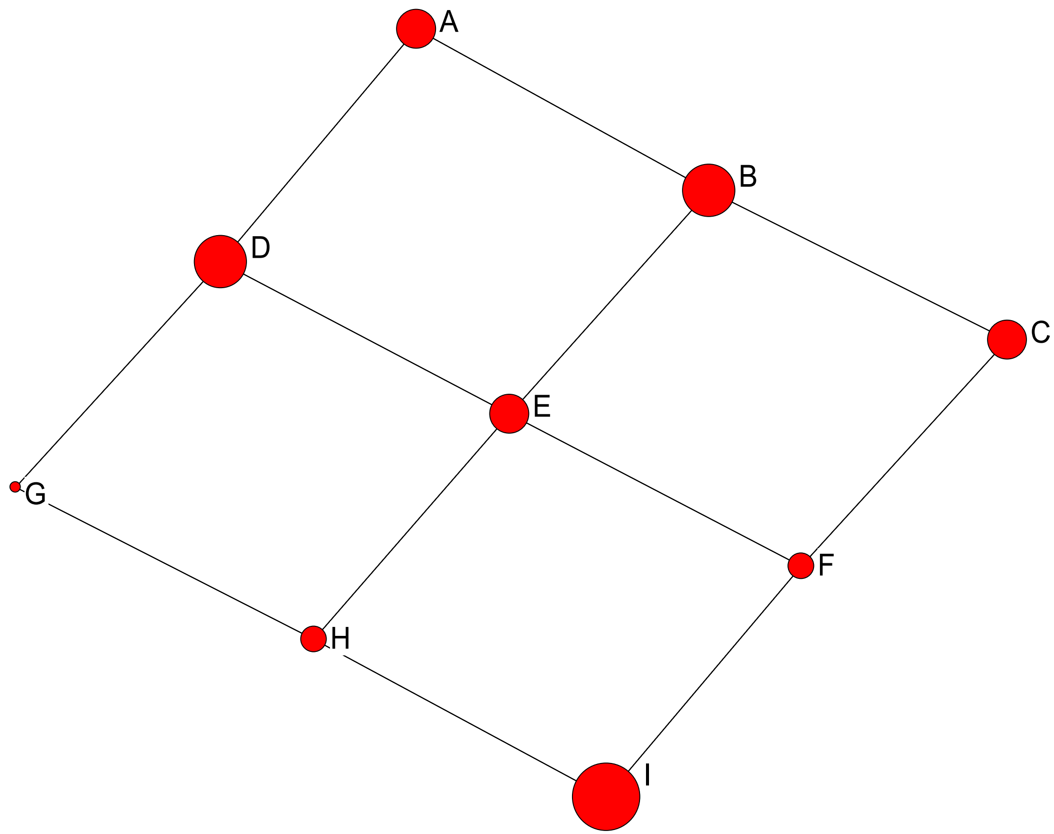


Node	Attrib
A	1
B	2
C	3
D	2
E	3
F	4
G	3
H	4
I	5

Geary's C: 0.333
Significance: 0.000

No Autocorrelation

Random pattern; (Geary's C = 1)

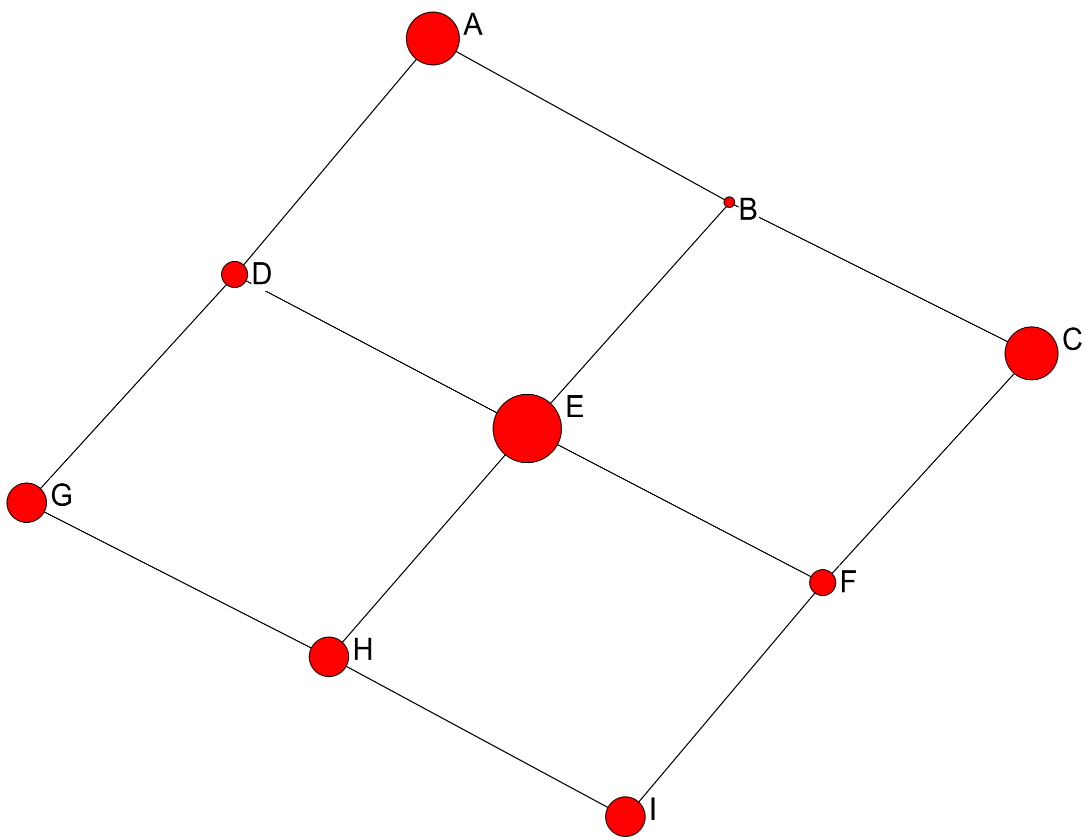


Node	Attrib
A	3
B	4
C	3
D	4
E	3
F	2
G	1
H	2
I	5

Geary's C: 1.000
Significance: 0.492

Negative Autocorrelation

(Dissimilars adjacent; Geary's C > 1)



Node	Attrib
A	4
B	1
C	4
D	2
E	5
F	2
G	3
H	3
I	3

Geary's C: 1.833
Significance: 0.000

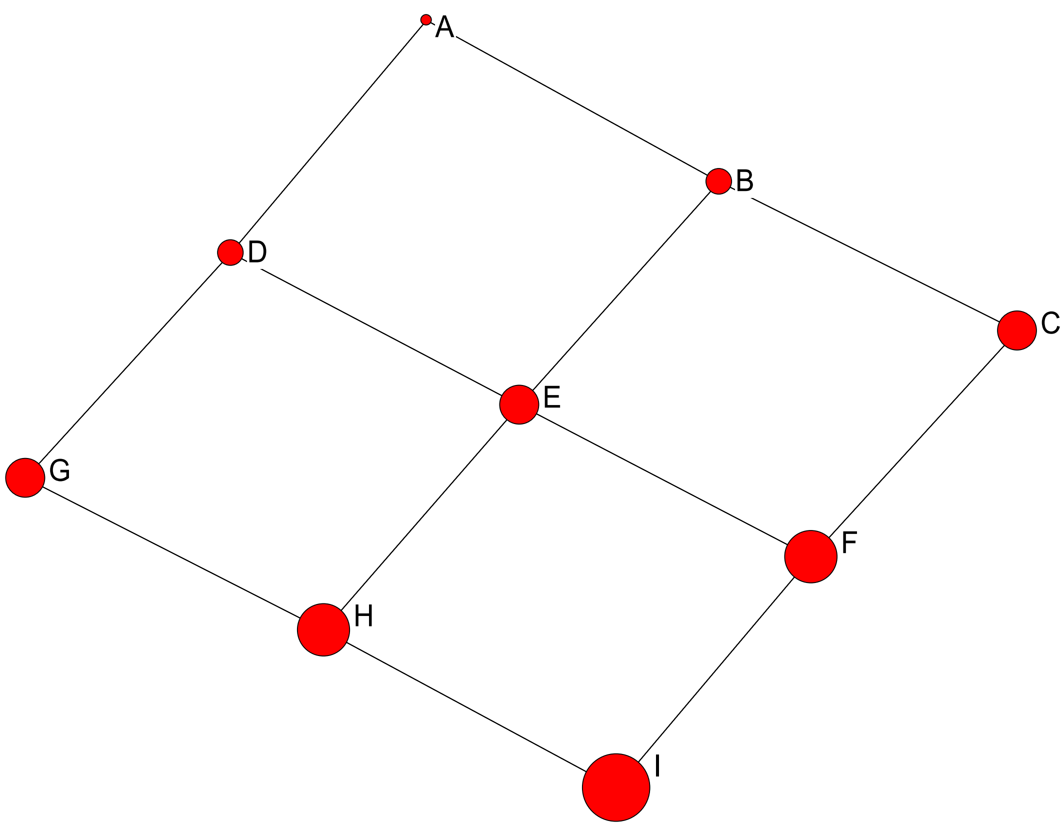
Moran's I

- Ranges between -1 and +1
- Expected value under independence is -1/(n-1)
- I → +1 when positive autocorrelation
- I → -1 when negative autocorrelation

$$I = n \frac{\sum_{i,j} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i,j} w_{ij} \sum_i (x_i - \bar{x})^2}$$

Positive Autocorrelation

(Similars adjacent; Moran's I > -0.125)

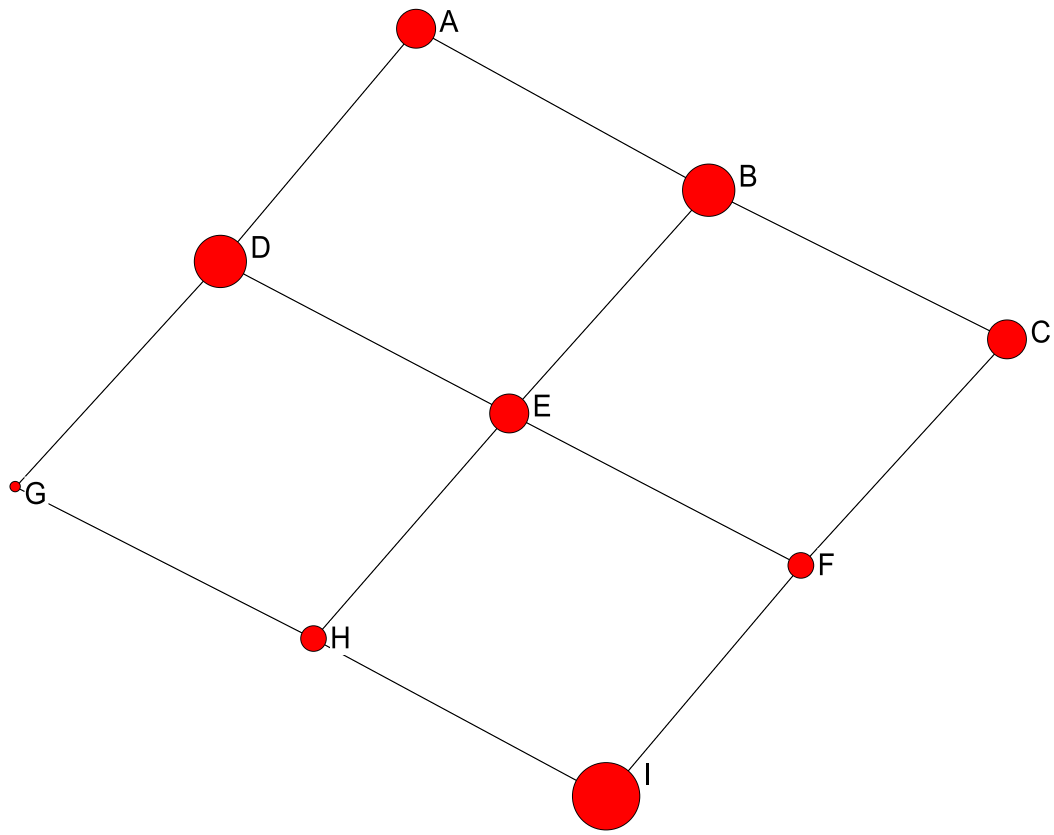


Node	Attrib
A	1
B	2
C	3
D	2
E	3
F	4
G	3
H	4
I	5

Moran's I: 0.500
Significance: 0.000

No Autocorrelation

Independence; (Moran's I ≈ -0.125)

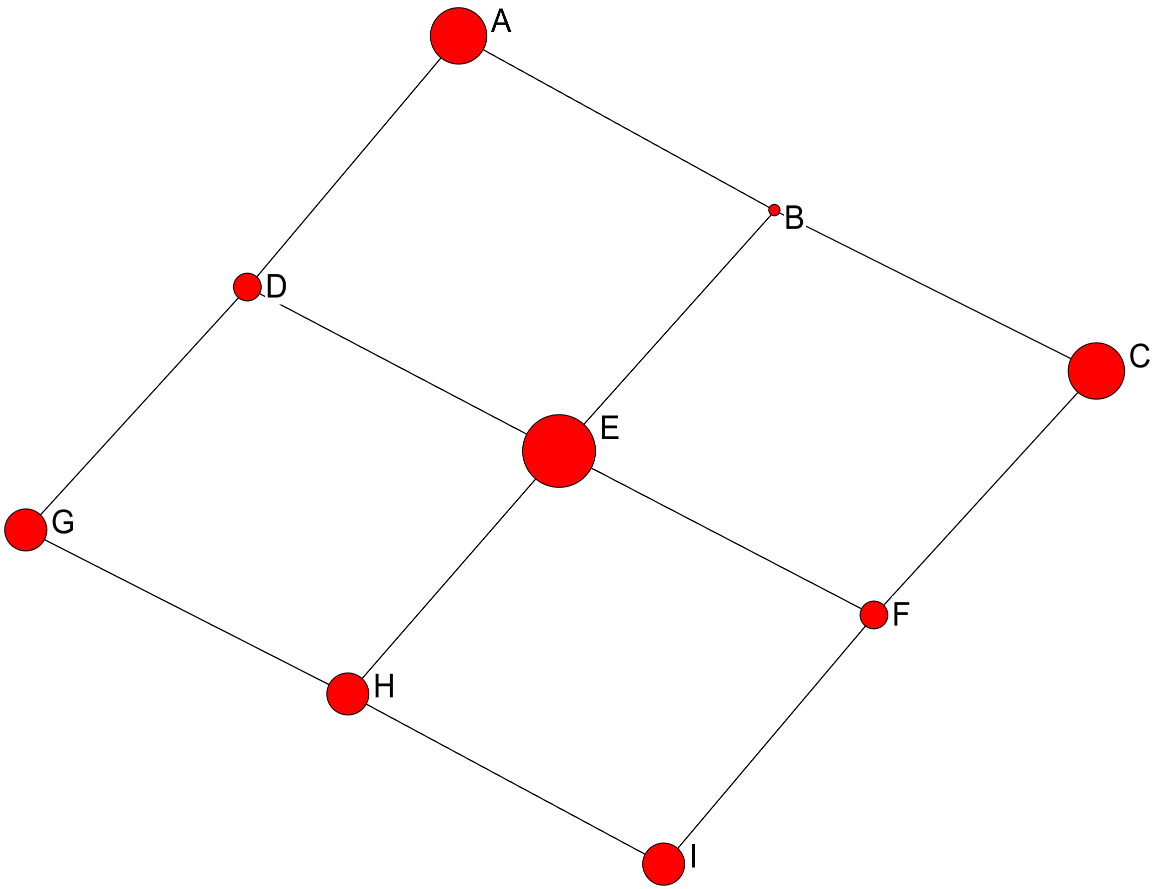


Node	Attrib
A	3
B	4
C	3
D	4
E	3
F	2
G	1
H	2
I	5

Moran's I: -0.250
Significance: 0.335

Negative Autocorrelation

(Dissimilars adjacent; Moran's I < -0.125)



Node	Attrib
A	4
B	1
C	4
D	2
E	5
F	2
G	3
H	3
I	3

Moran's I: -0.875
Significance: 0.000